An Open Online Dictionary for Endangered Uralic Languages

Hämäläinen, Mika

Lexical Computing CZ s.r.o.

2019

Hämäläinen , M & Rueter , J 2019 , An Open Online Dictionary for Endangered Uralic Languages . in I Kosem , T Zingano Kuhn , M Correia , J P Ferreira , M Jansen , I Pereira , J þÿ K a l l a s , M J a k u b í e k , S K r e k & C T i b e r i u s (e d s) , E l e c t r o n i c l e x i c o g r a : Proceedings of the eLex 2019 conference . Electronic lexicography in the 21st century , Lexical Computing CZ s.r.o. , Brno , pp. 819-830 , Electronic lexicography in the 21st century , Sintra , Portugal , 01/10/2019 .

http://hdl.handle.net/10138/305873

An Open Online Dictionary for Endangered Uralic Languages

Mika Hämäläinen, Jack Rueter

Department of Digital Humanities, University of Helsinki
E-mail: mika.hamalainen@helsinki.fi, jack.rueter@helsinki.fi

Abstract

We describe a MediaWiki-based online dictionary for endangered Uralic languages. The system makes it possible to synchronize edits done in XML-based dictionaries and edits done in the MediaWiki system. This makes it possible to integrate the system with the existing open-source Giellatekno infrastructure that provides and utilizes XML formatted dictionaries for use in a variety of NLP tasks. As our system provides an online dictionary, the XML-based dictionaries become available for a wider audience and the dictionary editing process can be crowdsourced for community engagement with a full integration to the existing XML dictionaries. We present how new automatically produced data is encoded and incorporated into our system in addition to our preliminary experiences with crowdsourcing.

Keywords: online dictionary; endangered languages; Uralic languages

1. Introduction

Open-source resources have been developed in the past for a number of endangered Uralic languages in the Giellatekno infrastructure (Moshagen et al., 2014). Giellatekno is the North Sami word for *language technology*, and work in the infrastructure at what today is known as the Norwegian Arctic University originally highlighted rule-based and finite-state descriptions of Sami languages in cooperation with the language communities. In addition to the Giellatekno research portion, a complementing implementational actor *Divvun* 'correction' has been established by the Sami Parliament for developing orthographic and morphological spellcheckers, keyboards, syntax checkers, machine translation, etc. Naturally, other Nordic languages are included in the infrastructure as well as minority languages of the Barents Sea and even larger Circum Polar Regions. The list of language projects amounts to over one hundred, with around 50 active projects. Some finite-state language descriptions now hosted date back to work in the early 1980s, while others are only now emerging.

Finite-state description with rule-based solutions at Giellatekno caters to languages with complex morphology. The philosophy at Giellatekno-Divvun includes multiple reuse of resources, i.e. by originally developing analysers for linguists, we are able to

produce almost simultaneously basic spellcheckers [1,2], and, at the same time, we can develop work with intelligent computer assisted language learning [3]. In late 2012 and early 2013 a project involving the development of online morphology-savvy dictionaries and click-in-text dictionaries was also spearheaded at Giellatekno for several well described languages, for example North Sami [4] and South Sami[5].

With the start of the Kone Foundation Language Programme, in Finland (2013–2017), it was decided that new minority language projects such as Livonian[6], Olonets-Karelian[7], Izhorian, Hill Mari[8], Erzya-Mordvin[9], Moksha-Mordvin, Komi-Zyrian[10] and Tundra Nenets[11] could readily be included among the online morphology-savvy dictionaries with spell relax mechanisms (see also Rueter, 2014). What was special about the newly introduced languages was that the online dictionary XML databases simultaneously served as the source for XSL transformation and transducer generation. Thus, basic information included in the XML files consisted of lemma, derivational stem, part-of-speech and specific inflectional type information, which was complemented by translations into Finnish and possibly other languages. Subsequent work with XML dictionaries has introduced additional languages, e.g. Skolt Sami[12], Udmurt, Komi-Permyak and Meadow Mari.

These XML resources featured in many of the Uralic language projects, however, are not easily available for people who are unfamiliar with technically advanced presentations, as they are provided in source code format.

We present a MediaWiki-based multilingual online dictionary for endangered Uralic languages. The dictionary not only makes the lexicographic resources available for ordinary users, but it makes dictionary editing possible in a crowd-sourced fashion with an XMLMediaWiki synchronization (Hämäläinen & Rueter, 2018). This means that any edits made in the original XML files in the Giellatekno infrastructure will be synchronized to the online dictionary, and vice-versa.

---

1 http://divvun.no/

2 http://divvun.org/

3 http://oahpa.no/davvi/

4 https://sanit.oahpa.no/

5 https://baakoeh.oahpa.no/

6 http://sonad.oahpa.no/

7 http://sanat.oahpa.no

8 http://muter.oahpa.no/

9 http://valks.oahpa.no

10 http://kyv.oahpa.no/

11 http://vada.oahpa.no

12 http://saan.oahpa.no

The lexicographic entries in our online dictionary have been automatically enhanced with a multitude of Semantic MediaWiki tags. In the past, Semantic MediaWiki has been shown to be a viable way of integrating semantic web compatible information with an online dictionary (Laxström & Kanner, 2015). Our online dictionary also provides an API access to its resources. Over the API, lexicographic entries can be retrieved in JSON format and the FST transducers can be used both for morphological analysis and generation.

In this paper, we provide insight on the functionalities of our MediaWiki-based online dictionary system. Furthermore, we describe how lexicographic information newly obtained by using language technology approaches is incorporated into the online dictionary.

Currently, we support 13 endangered Uralic languages such as Skolt Sami, Komi-Zyrian, Udmurt and Erzya. We have initially experimented with crowd-sourcing for Skolt Sami and Erzya with positive results.

2. Related work

In the modern era, developing accessible and easy to use dictionaries for endangered languages has become one of the important research interests in language documentation and revitalization. Some of the work focuses more on building a new dictionary out of scratch, whereas others focus on making already existing paper dictionaries accessible for a wider audience in a much more modern fashion. In this section of the paper, we describe some of the contemporary work on this topic.

Work with endangered languages in North America has shown that the language novice must be provided for. The communities are small, and unfamiliarity with lexicographic tradition can easily be detrimental to the novice's language learning experience. The new language learner cannot be expected to know where a dictionary entry lies nor automatically adopt the normative orthography. When the language user either lacks the keyboard or the knowledge to spell correctly, spell relax strategies can be implemented in online and mobile morphology-savvy solutions. Morphologic awareness and spell relax are used in catering to the Tsimshianic and Salish novice in dictionary use and language technology (Littell et al., 2017). On an entirely separate front, work has also been done to provide the St. Lawrence Island Yupik community with unhindered access to language materials online. This, once again, has been accomplished using a morphologically aware dictionary. In this separate rendering of the same kind of system, however, a strategy of multiple input methods catering to different writing systems (Hunt et al., 2019) has been introduced. The work here is tailored, and a strong tie is maintained between a language and its community. These endangered languages fall into the category of low-resourced languages.

'Low-resourced language', however, is a term used for almost any language with a lower internet presence than English. In (Nasution et al., 2018), in contrast, the Malaysian languages dealt with are relatively small in comparison to the majority languages encompassing them. The approach is to address a group of closely related languages simultaneously – an underlying multilingual or language-independent infrastructure. Pivot languages are used as means of enriching bilingual lexical resources. The authors discuss drawing upon bilingual dictionary input, and the difficulty of selecting the right bilingual dictionaries to start from.

One part of the strategy is to use cognates found through pivot-languages for locating translation candidates. Cognates are subsequently paired with multiple synonyms, and these synonym continua are established in many-to-many translation blocks. This is one of the places where native speaker editors are employed in the evaluation of automatically generated much needed lexica. Since the focus is on a larger language populations, outlines are made of actual expenses incurred in editing bilingual lexical resources, i.e. expenditures based on 10 and 30-second increments in an eight-hour day.

Low resource endangered languages do not necessarily have the native speaker-editor population to draw upon. Therefore, language-independent approaches are merited even here.

3. The MediaWiki-based dictionary

The main motivation behind the use of MediaWiki is to make the Giellateknko XML dictionaries authored for a multitude of endangered Uralic languages available for the general public. This is done in a synchronized way so that edits done in both the XMLs and the MediaWiki can be synchronized. This will ensure the availability of the latest version of the data for all users.

Uralic languages are known for their highly inflectional morphology. This makes the use of traditional dictionaries difficult, as a language learner will have to successfully inflect a word form he has encountered in a text to its lemma form in order to find it in a dictionary.

To alleviate this problem, our online dictionary includes finite-state morphological analysers (cf. Beesley & Karttunen, 2003) that will lemmatize the user input before querying the lexicographic database. In this way, the user can find the lemma and its translations even when it comes to morphologically complex word forms. These analysers are generated from the XMLs that can be edited in the MediaWiki system (cf. Rueter & Hämäläinen, 2017).

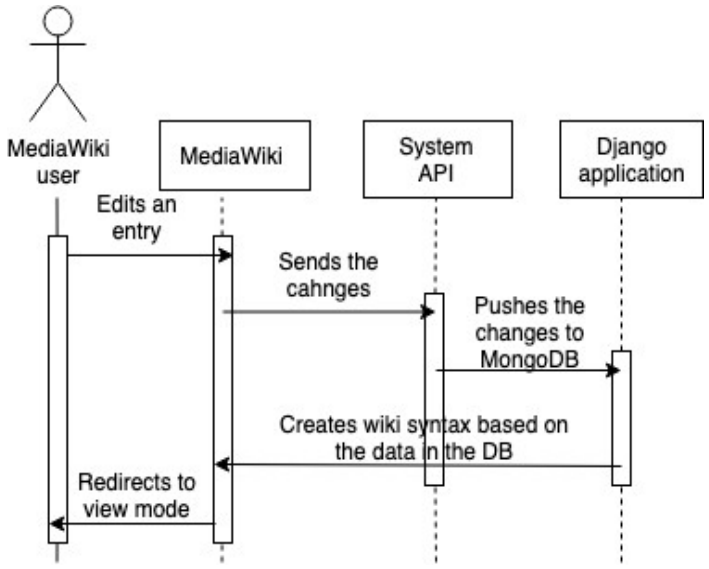

Figure 1: A diagram showing an edit on the MediaWiki side.

The synchronization of editing has been done in such a way that the up-to-date data is available for both the people working with the XMLs and on the MediaWiki. This is important as technically savvy people find XML-based editing more powerful whereas non-technical people would have problems working with the markup, where even adding a wrong character might render the whole XML syntax invalid. Figure 1 shows the process from the point of view of the person doing edits on the MediaWiki. Whenever the user is done with editing an entry in the dictionary, a Django-based synchronization system is informed. The Django system keeps an up-to-date backup in JSON format of all the entries in the dictionary. The edited entry is sent by a MediaWiki extension as JSON to the Django-based system, which updates its own database with the updated entry and re-formats the data in MediaWiki syntax to store it in the MediaWiki dictionary for visualization to the dictionary users.

Editing the XMLs is a slightly more complicated process, as shown in Figure 2. We have decided to build the XML editing on top of Git as it provides versioning and it makes it possible to compare the different versions and resolve potential conflicts in an easy to use fashion, especially due to the availability of a myriad of Git tools with a graphical user interface. The process starts by the lexicographer using a custom Git script to pull the latest version of the XML from the Django system running on the server of the MediaWiki system.

Once the lexicographer is done with the edits of the XMLs, he can push the changes to the master branch of the GitHub repository. This will initiate a pull on the MediaWiki server and the Django-based system starts a background process to first update its own internal database with the changes in the XML files, and then generate and update MediaWiki syntax for the updated entries.

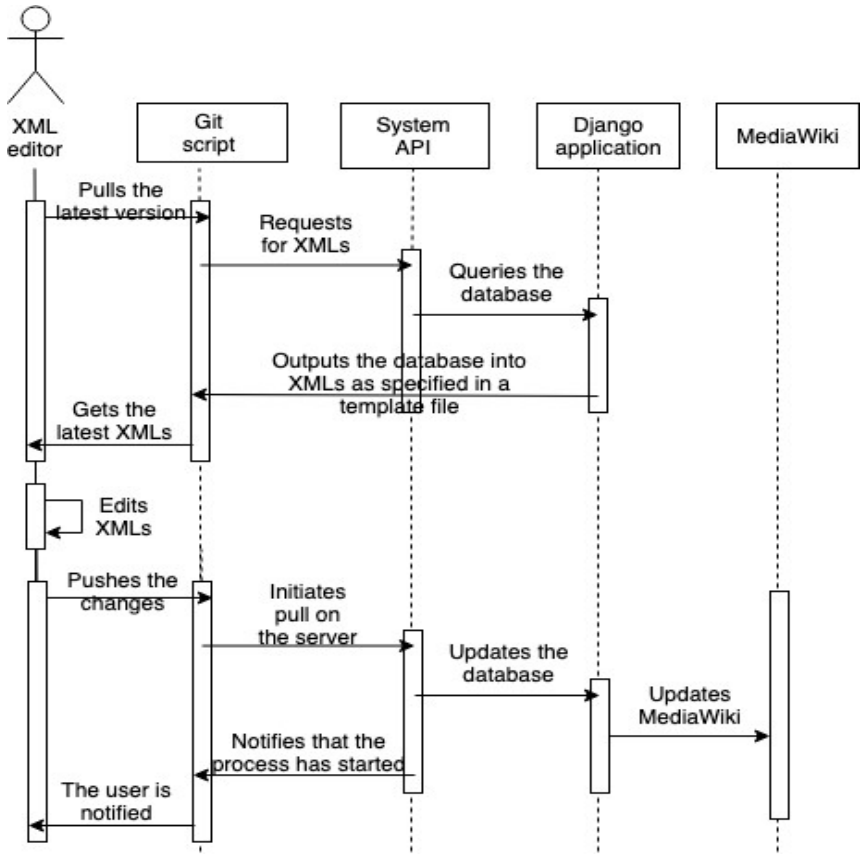

Figure 2: A diagram showing an edit on the XML side.

4. Representing the new information

This section of the paper is dedicated to describing how the data obtained by automatic language-technology methods for Uralic languages has been incorporated to our MediaWiki-based online dictionary system. Making the new data available on a system that also serves for non-academic usage is important not only for revitalization of the endangered Uralic languages, but also for community involvement.

Hämäläinen et al. (2018) presents work on combining dictionaries automatically for Skolt Sami, Erzya, Moksha and Komi-Zyrian based on the XML dictionaries also available on our MediaWiki dictionary. As all of the dictionaries are multilingual, meaning that every entry in a minority language has translations into multiple majority languages (most typically Finnish, English and Russian), it is possible to combine translation entries for all of the four minority languages. This is based on two assumptions, firstly the XML structure has meaning groups, which means that translations are grouped by senses, and secondly if a meaning group has translations into two different languages, the languages will make a semantic distinction and therefore translations that do not refer to the same meaning are not combined.

In practice, the approach takes an entry in Skolt Sami, such as *bliin*, which has translations into Finnish *ohukainen* and *blini* and in English *pancake* and compares it to an entry in Komi-Zyrian, which in addition to the same translations as in the Skolt

Entry, also has the synonyms *räiskäle* in Finnish and *crepe* in English. As there is an overlap between the entries, the method extends the Skolt Sami entry with the additional synonyms from the Komi-Zyrian entry.

In order to incorporate these results into our MediaWiki dictionary, it is important to introduce a new attribute to the XML structure, namely an ID for each individual meaning group. When the meaning groups can be identified, the linking of the dictionary entries can be done on the system level. Currently, a hand-curated set of the automatic results presented in Hämäläinen et al. (2018) are included in our online dictionary. In the future, their approach could be included in a dynamic fashion in our system so that whenever a new entry is added on the MediaWiki platform, a set of possible translations together with links to meaning groups in other languages could be brought as suggestions to the dictionary editor.

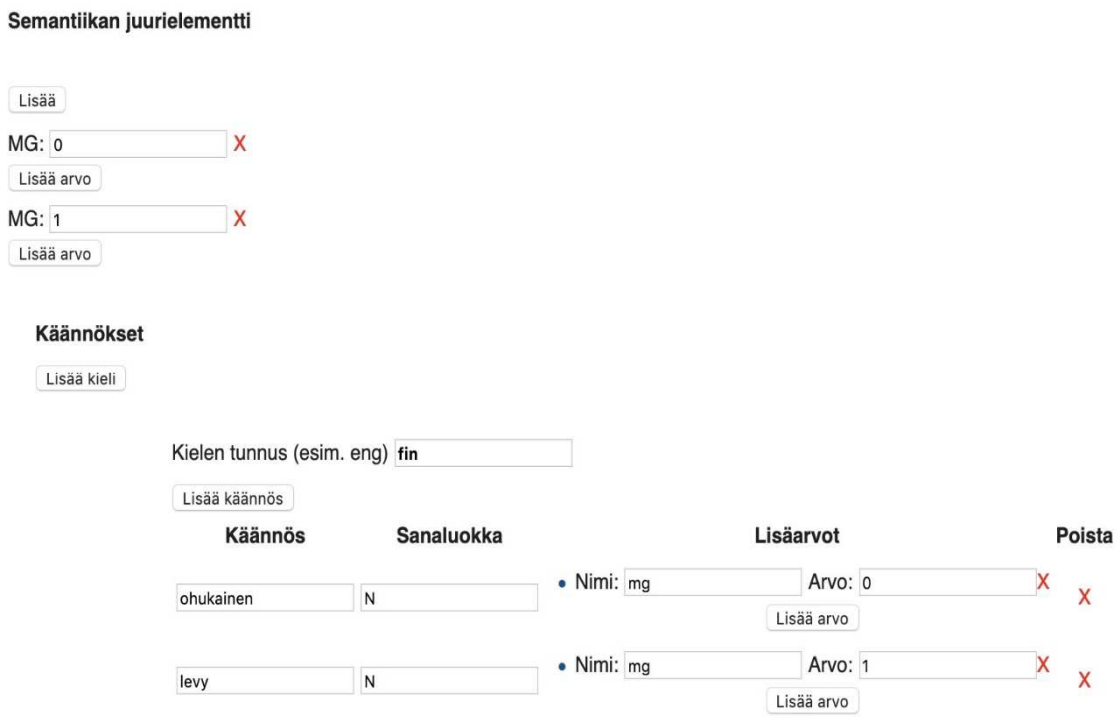

Figure 3: Meaning groups in the MediaWiki edit form.

Meaning groups (MGs) have editable locally unique IDs in the edit form of MediaWiki, as seen in Figure 3. Meaning groups can be added as needed. Translations in different languages are grouped together when the dictionary data is visualized for the user based on the meaning group IDs.

SemUr and SemFi (Hämäläinen, 2018) are automatically extracted semantic databases for Skolt Sami, Erzya, Moksha, Komi-Zyrian and Finnish. These databases represent corpus frequencies of co-occurrences of two words given a syntactic relation. Through this data it is possible, for instance, to see which words can act as a subject or object for a given verb. This can be a useful resource for a lexicographer especially as it reveals

information about polysemy, not to mention the number of links it introduces in between the different dictionary entries.

The graph like relation structure calls for a different visualization strategy to what is commonly used in MediaWiki. Therefore, we create our own MediaWiki extension that can be used to visualize and browse the semantic databases. This visualization can be accessed from a dictionary entry on the MediaWiki.

Figure 4 shows the interface incorporated in our MediaWiki-based dictionary for browsing the semantic data. In the example, the adjective modifiers and verbs with the subject relation are shown for the Finnish word *kirves* 'axe'. The interface gives the possibility to focus on related words of a certain part-of-speech or syntactic relation.

Recent work using neural networks to extend cognate relations for Skolt Sami and North Sami (Hämäläinen & Rueter, 2019) is an important data point for lexicographic work. Cognates from closely related languages can further be used in a multitude of language technology applications. Cognate relations are introduced to our online dictionary by linking words sharing a cognate relation to each other. This way, a dictionary user can move from one entry to its cognate easily. The same linking functionality is also used to link compound words with their constituents.

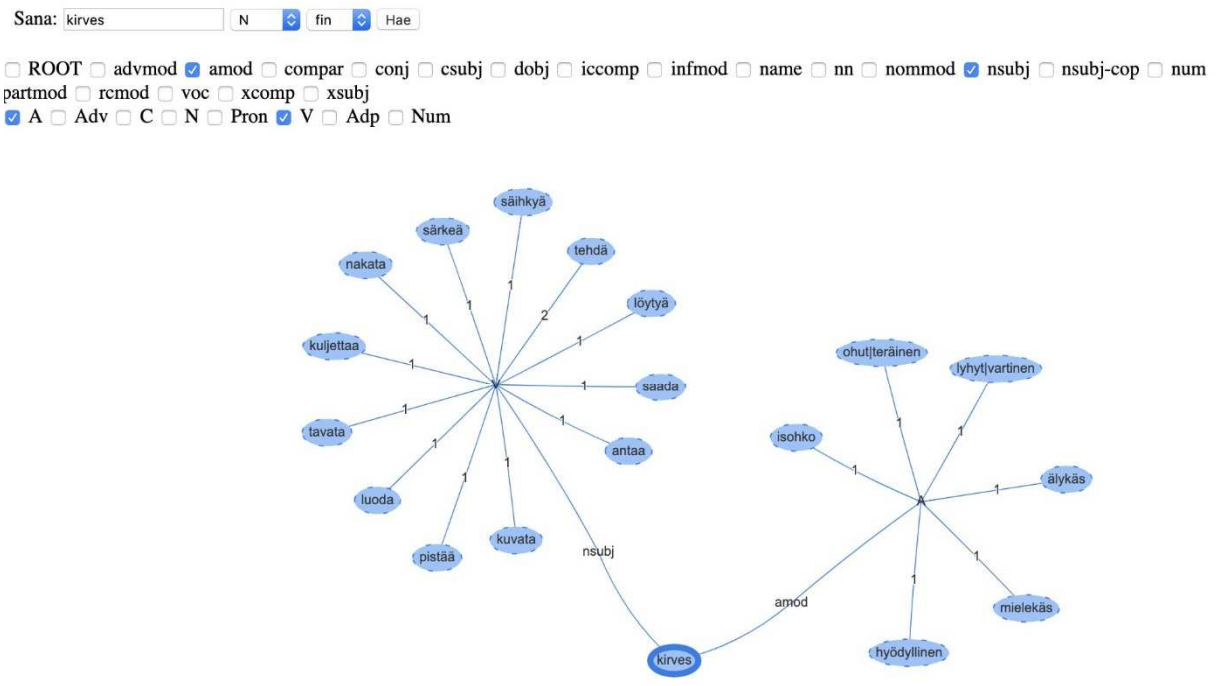

Figure 4: Interface for browsing semantic data.

Figure 5: Cognate view in the dictionary.

Cognates can be viewed by clicking on a button titled *Näytä etymologia* 'show etymology', as seen in Figure 5. Information is shown about the cognate word together with a link to its entry in the other language.

All of this new information introduced into the system has been made available for programmatic access over the custom API of the MediaWiki dictionary. The access to this API has been integrated into UralicNLP Hämäläinen (2019), which is an open-source Python library for processing endangered Uralic languages.

5. Crowd-sourcing

Our initial experiments in crowd-sourcing have been limited to a small number of people due to the fact that the communities speaking the endangered languages in question are not as big as they are in the case of majority languages. Nevertheless, crowd-sourcing serves for the purpose of exposing the XML structured dictionaries to non-technical linguists and community members.

Work with the Skolt Sami, Erzya and Komi-Zyrian language communities has included actual editing of MediaWiki materials that have directly augmented the dictionary database and hence enhanced the materials and tools available on the parallel Giellatekno infrastructure. During the summer of 2017, one work involving community linguists added much needed verbal derivation content in addition to example sentences from language archive materials at the Giellagas Institute in Oulu, Finland. In this two-month trial, conflicts between MediaWiki editors and XML editors were resolved. Additional input parameters that were found necessary were incorporated into the infrastructure to allow for sound-to-text alignment of archive materials in future work with Skolt materials, i.e. this was ground-breaking with regard to future work with other languages as well.

A second encounter with community collaboration was organized at the end of 2017. This time around, native and virtually native speakers were asked to evaluate automatically aligned concept translations. The alignments consisted of one source-language word with translations into several target-language words. The alignment had been facilitated using two pivot languages. In this way, new translations were shared

between dictionaries for the source languages Skolt Sami, Erzya, Moksha and Komi-Zyrian. Translation languages included English, Finnish, French, German, North Sami, Norwegian, and Russian, as well as some other minority languages. The task consisted of (i) accepting, (ii) not accepting, and if not accepting (iii) noting. Although the nature of the task was relatively straightforward, finding native speakers with adequate knowledge in three or more languages was a problem, but not entirely unsurpassable.

Crowd-sourcing introduces issues of access and tools in general. Work with language communities lacking active representatives in the Finnish academic community introduces a need for issuing non-academic usernames and access. This required the system to be moved away from using Haka credentials, which is a nation-wide authentication system for academic institutions in Finland. Levels of access must then be established that, on the one hand, allow access to language community activists and researchers and, on the other hand, ensure the integrity of the open-source multilingual lexical data synchronously maintained in Tromsø, Norway and Helsinki, Finland. Once access has been established, there is a need to maintain quality control of the data, i.e. one source of problems is that Skolt Sami has several Latin characters available only on a few open-source keyboards, the same applies to Komi-Zyrian and the Mari languages, which have letters from outside the Russian Cyrillic alphabet – should there be a virtual keyboard available.

6. Discussion and future work

Our online dictionary system represents a big leap towards the correct direction in making language resources available both for regular dictionary users and for more technically oriented users through the open API. However, as indicated by our crowd-sourcing experiments, some additional care has to go into streamlining the usability of the dictionary editing. Currently, the edit form reveals a myriad of detailed information such as continuation lexicon and stem group, which might be overwhelming for an average language speaker. This calls for more user-centric usability testing to be conducted in the future.

The combined meaning groups from Hämäläinen et al. (2018) have been introduced into the system in a static fashion. However, their method could, in the future, be integrated into our system in a more dynamic way. In practice, this would mean that a dictionary editor adding a new entry for any language in the system would get recommendations for other candidate translations to choose from. This could speed up the process of conducting lexicographic work with endangered languages.

More active engagement of the community members is needed in the future. The first step to make contributing to the dictionary as easy as possible would be localization of the interfaces used. First and foremost to Russian, as a vast majority of the endangered Uralic languages are spoken in Russia, but also localization to all the supported endangered languages.

7. References

Beesley, K. R. & Karttunen, L. (2003). *Finite-State Morphology*. Stanford, CA: CSLI Publications, pp. 451–454.

Hunt, B., Chen, E., Schreiner, S. L. & Schwartz, L. (2019). Community lexical access for an endangered polysynthetic language: An electronic dictionary for St. Lawrence Island Yupik. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 122–126. https://www.aclweb.org/anthology/N19-4021.

Hämäläinen, M. (2018). Extracting a Semantic Database with Syntactic Relations for Finnish to Boost Resources for Endangered Uralic Languages. In *Proceedings of the Logic and Engineering of Natural Language Semantics 15 (LENLS15)*.

Hämäläinen, M. (2019). UralicNLP: An NLP Library for Uralic Languages. *Journal of Open Source Software*, 4(37), p. 1345.

Hämäläinen, M. & Rueter, J. (2018). Advances in Synchronized XML-MediaWiki Dictionary Development in the Context of Endangered Uralic Languages. In *Proceedings of the Eighteenth EURALEX International Congress*. pp. 967–978.

Hämäläinen, M. & Rueter, J. (2019). Finding Sami Cognates with a Character-Based NMT Approach. In *Proceedings of the Workshop on Computational Methods for Endangered Languages*, volume 1.

Hämäläinen, M., Tarvainen, L. L. & Rueter, J. (2018). Combining Concepts and Their Translations from Structured Dictionaries of Uralic Minority Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA).

Laxström, N. & Kanner, A. (2015). Multilingual Semantic MediaWiki for Finno-Ugric dictionaries. In *Septentrio Conference Series*, volume 2, pp. 75–86.

Littell, P., Pine, A. & Davis, H. (2017). Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Honolulu: Association for Computational Linguistics, pp. 141–150. https://www.aclweb.org/anthology/W17-0119.

Moshagen, S., Rueter, J., Pirinen, T., Trosterud, T. & Tyers, F. M. (2014). Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. In *The LREC 2014 Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era"*, pp. 71–77.

Nasution, A.H., Murakami, Y. & Ishida, T. (2018). Designing a Collaborative Process to Create Bilingual Dictionaries of Indonesian Ethnic Languages. In *Proceedings of the 11th Language Resources and Evaluation Conference*. Miyazaki, Japan: European Language Resource Association. https://www.aclweb.org/anthology/L18-1536.

Rueter, J. (2014). The Livonian-Estonian-Latvian Dictionary as a threshold to the era of language technological applications. *Eesti ja soome-ugri keeleteaduse ajakiri.*

Journal of Estonian and Finno-Ugric Linguistics, 5, p. 251.

Rueter, J. & Hämäläinen, M. (2017). Synchronized MediaWiki Based Analyzer Dictionary Development. In *Proceedings of the Third Workshop on Computational Linguistics for Uralic Languages*, pp. 1–7.